A Bi-Dimensional User Profile to Discover Unpopular Web Sources

Romain Noel
LITIS, Normandie Université
INSA de ROUEN
Airbus Defense & Space
Val-de-Reuil, France
romain.noel@cassidian.com

Nicolas Malandain
LITIS, Normandie Université
INSA de ROUEN
St Etienne du Rouvray, France
nicolas.malandain@insa-rouen.fr

Alexandre Pauchet
LITIS, Normandie Université
INSA de ROUEN
St Etienne du Rouvray, France
alexandre.pauchet@insa-rouen.fr

Laurent Vercouter
LITIS, Normandie Université
INSA de ROUEN
St Etienne du Rouvray, France
laurent.vercouter@insa-rouen.fr

Bruno Grilheres
Airbus Defence & Space
Val-de-Reuil, France
bruno.grilheres@cassidian.com

Stephan Brunessaux
Airbus Defence & Space
Val-de-Reuil, France
stephan.brunessaux@cassidian.com

ABSTRACT

The discovery of new sources of information on a given topic is a prominent problem for Experts in Intelligence Analysis (EIA) who cope with the search of pages on specific and sensitive topics. Their information needs are difficult to express with queries and pages with sensitive content are difficult to find with traditional search engines as they are usually poorly indexed. We propose a double vector to model EIA's information needs, composed of DBpedia resources [2] and keywords, both extracted from Web pages provided by the user. We also introduce a new similarity measure that is used in a Web source discovery system called DOWSER. DOWSER aims at providing users with new sources of information related to their needs without considering the popularity of a page. A series of experiments provides an empirical evaluation of the whole system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

1. INTRODUCTION

The explosive growth of the Web has resulted in a huge amount of information available on Internet. Finding relevant sources has become a complex task. Experts in Intelligence Analysis (EIA) often explore the Web to collect information on specific and sensitive topics such as Web sites selling illegal pharmaceutical products, Jihadist blogs, terrorist forums and so on. Their information needs are therefore to discover new information sources on search topics.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742139 .

The search activity of EIA has some specific characteristics that make traditional Information Retrieval (IR) tools unsuitable. First of all, EIA find difficult to express their needs using traditional queries, as the vocabulary of sensitive pages quickly evolve. For instance, new drug names, synonyms of the same molecule, often appear, and thus EIA need to discover new sensitive pages to update their vocabulary. Then, sources containing sensitive content are usually poorly indexed in traditional web search engines because of their lack of popularity. Such information sources can also be not indexed at all in order to stay unreachable by lambda users or because search engines deprecate their content. Finally, EIA have to combine broad search and deep search to explore sources on an identified relevant topic but also to consider, and sometimes discover, new related topics.

In this article, we introduce an original approach of user profile modelling to address the problem of sensitive need representation for EIA. Instead of queries, we propose to describe a user's information needs with a double vector of DBpedia resources [2] and keywords to cover respectively the thematic and the specific aspects of her information need. The user profile is constructed semi-automatically to avoid EIA to use their own list of terms. To tackle the problem of poorly indexed web sites, we exploit our own focused crawler called DOWSER (Discovery Of Web Sources Evaluating Relevance) [12] that integrates a new similarity measure to index pages regardless of their popularity.

Our approach provides the following main contributions: (i) a semi-automatically constructed user profile based on DBpedia concepts and keywords both used by DOWSER; (ii) an approach for relevance calculation based on this profile; and (iii) an automatic ranking process to provide relevant sources of information to the user. In section 2, we compare our approach to existing works. In section 3, the user profile representation and our similarity measure are described. A user experiment and the results obtained are presented in section 4. Finally, we conclude by discussing possible extensions in section 5.

2. RELATED WORK

Most of the existing web search engines work on a set of indexed pages collected by a crawler. Usually, documents are indexed independently of any information needs and a query is then used to retrieve pages of interest but search engines rank them mostly according to their popularity and not only depending on their suitability to user needs. Recent works propose two solution to overcome those problems: personalized IR systems and focused crawling.

2.1 Personalized search systems

A personalized search is based on a user profile that encodes the information needs, exploited in a search engine.

2.1.1 Modelling the user profile

The construction of a user profile can be explicit or implicit. An implicit approach aims to automatically capture the user interests. The most widely used techniques consists in extracting information from the user's search history [9, 19]. On the other hand, an explicit construction of a user profile requires from the user to interact actively with an adequate interface [20]. The user can communicate her preferences and interests to the system, by providing a set of relevant documents or by compiling questionnaires. The weakness of this approach is that the user profile construction is based on her ability to express her information needs.

The representations of a user profile are usually based on vector models [21] or bag of weighted-keywords [18]. Nowadays, approaches also integrate external knowledge to improve the quality of the user profile representation. Using an ontology, the representation of a user profile is more structured. An ontology formally represents knowledge as a set of concepts to determine the search context and the user interests using predefined semantic resources (e.g., DBPedia[1] [2]). These approaches can represent the user interests with concepts automatically extracted from the user's documents.

2.1.2 Exploiting the user profile

Several approaches have been proposed to exploit a user profile into an IR system, before or after the search engine process. Therefore, the set of approaches can be classified into two main categories [11]: (a) The user profile is used during a distinct re-ranking step to increase the precision of the ordering process. (b) The representation of the information needs is affected by the user profile as user's queries are modified by adding or changing keywords.

These two approaches improve the efficiency of search engines via (a) personalized queries and/or (b) re-ranked results. However, personalized search engines work on a set of Web pages indexed from the whole Web without considering the user needs.

2.2 Focused crawling

Traditional crawlers explore Web pages from a URL queue and convert them into plain text to extract the contained links. Those links are added to the URL queue in order to crawl other Web pages. According to a set of acceptance rules, collected Web pages are then indexed. Traditional crawlers are based on graph algorithms, such as breadth-first or depth-first traversal, to explore the Web.

[1]http://dbpedia.org/About

On the contrary, focused crawling [1, 3] aims to improve directly the crawling phase by collecting only pages related to the user needs. A focused crawling system exploits additional information to predict if the page is relevant. For example, focused crawlers can reject pages [4] using anchor text of source URL [5]. Figure 1 presents a crawler focusing its exploration by collecting pages with the highest priority.

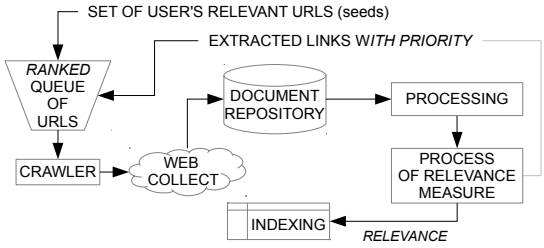

Figure 1: Focused crawler

According to the topical locality phenomenon [5], Web pages connected together should concern similar topics. Focused crawlers are based on this phenomenon and crawl clusters of pages each time they find an interesting page. Therefore, the two famous algorithms PageRank [13] and HITS [8], based on the structure of links, are really effective. They are often used to assign a hypertextual-based rank to seeds[2].

2.3 Limits of existing systems for EIA

Finding sensitive and specific Web pages is a task that needs some adjustments to existing approaches from IR. Personalized IR tools improve the representation of the user's information needs [11] but exploiting a user profile within a search engine is unsuitable when the needed pages are unpopular or part of the Deep Web. A solution is to use a personalized focused crawler, but contrary to Chakrabarti et al. [3] and Bergmark et al. [1], the popularity of a page should not be used to rank results.

Furthermore, using only keywords or only concepts to index pages [1, 15] may be a limitation if the content of needed pages is too specific. We therefore propose to use both keywords to model accurately the user needs and concepts to provide a wider and thematic coverage of the user needs.

Finally, existing works have to be adapted to a "more-like-this" approach in order to find unpopular Web pages with specific content. This correspond more precisely to the EIA's task than defining precise requests.

3. DISCOVERING UNPOPULAR SOURCES WITH SENSITIVE CONTENT

The aim of our approach is to automatically discover relevant sources that match with a user's information need and that could not be easily found with classic search engines. We propose a system that exploits a user profile composed of a double vector to guide a focused crawler. A similarity measure is computed to assess the relevance of collected pages according to the user profile.

[2]List of URLs to visit

3.1 A user profile based on provided seeds

Providing a request to express a thematic information need can be difficult for EIA. Usually, EIA already have a limited set of URLs on a specific and sensitive topic and need to find new relevant sources on the same topic. In our approach, the user is therefore helped in the construction of her profile by semi-automatically extracting relevant terms from a set of Web pages she has provided.

To represent a user need of information as a vector of terms, both keywords and concepts can be used. Concepts have the advantage not only to disambiguate extracted information but also to offer the possible extension of a user's need using subsumption hierarchy. However, the extraction of concepts is always limited to existing resources in a domain ontology and therefore important terms can be omitted if only a conceptual approach is used. On the contrary, keywords can describe specific and precise information needs. We therefore propose to use a bi-dimensional user profile, composed of a double vector of instances of concept and of keywords to offer a representation respectively thematic and specific of the user need of information.

On term of thematic representation, we use the DBpedia Spotlight[3] service to extract DBpedia Resources [2] for each Web page provided by the user. A weighted-concept vector is constructed by considering the frequency of occurrence of each DBpedia Resource in the set of pages [10]. To construct the keyword representation, the system is based on Apache Lucene[4] to extract relevant keywords thanks to the NP Chunker method [14]. The extracted keywords are weighted according to the $TF.IDF$ measure.

Our method of user profile construction is semi-automatic. After a first step of automatic extraction of ten weighted keywords and ten weighted instances of concepts, a manual selection of terms is needed to validate the user's interests. These two vectors represent the user's information needs that are processed next by our similarity measure during the discovery task of new relevant sources.

3.2 A Similarity measure combining thematic and keywords aspects

A focused crawler has to focus its collect on URLs linked to pages of interest. The relevance of a collected page is assessed according to the user need using a similarity measure. First of all, DBpedia Spotlight and the NP Chunker process a collected page similarly than during the construction of the user profile, in order to obtain a vector of weighted DBpedia Resources and a vector of weighted keywords representing the topics of the collected page. Then, the thematic relevance of a page is calculated as the result of a similarity measure between the vector of concepts representing the collected page and the vector of concepts from the user profile (the thematic part). It is based on the approach of Milne and Witten [22], modified by Saint Requier [17]: the first category (i.e. the DBPedia hierarchy ranked by topic) shared by two DBpedia Resources sets their semantic proximity. This measure is adequate to our problem because categories can be exploited to have a thematic representation of the user needs. A more accurate measure between concepts (e.g. [7, 16]) is not necessary since our goal is to categorize and disambiguate the thematic need.

[3]http://spotlight.dbpedia.org/

[4]http://lucene.apache.org/core/

Let $\vec{P}_C$ be the thematic representation of a user profile composed of a vector of N concepts $P_C(i)$ having a weight associated $W_{P(i)}$. Similarly, let $\vec{D}_C$ be the concept vector of the collected document D. Finally, let $cat(c)$ be a function that returns the set of categories of a concept c (each category i returned by this function is noted $cat_i(c)$). The thematic similarity measure is computed as follows:

$$Sim_C(\vec{P}_C, \vec{D}_C) = \frac{1}{|\vec{D}_C|} * \sum_{r=1}^{|\vec{D}_C|} \sum_{i=1}^{|\vec{P}_C|} \frac{W_{P_C(i)}}{Max(|cat(D_C(r))|, |cat(P_C(i))|)} * \sum_{j=1}^{|cat(D_C(r))|} \sum_{k=1}^{|cat(P_C(i))|} Sim(cat_j(D_C(r)), cat_k(P_C(i))) \quad (1)$$

where:

$$Sim(cat_j(r), cat_k(P_C(i)) = \begin{cases} 1 & \text{if } cat_j(r) = cat_k(P_C(i)) \\ 0 & \text{otherwise} \end{cases}$$

and therefore

$$Sim_C(\vec{P}_C, \vec{D}_C) \in \mathbb{Z}, 0 \leqslant Sim_C(\vec{P}_C, \vec{D}_C) \leqslant 1$$

Contrary to concepts, keywords are used to represent a specific need of information with terms that can be missing in a taxonomy of concepts. The cosine similarity is used on the two vectors of keywords to compute the specific relevance of a collected page:

$$Sim_K(\vec{P}_K, \vec{D}_K) = \cos(\vec{P}_K, \vec{D}_K).$$

where $\vec{P}_K$ and $\vec{D}_K$ are respectively the keyword vector of the user profile P and the keyword vector of the collected document D, with

$$Sim_K(\vec{P}_K, \vec{D}_K) \in \mathbb{Z}, 0 \leqslant Sim_K(\vec{P}_K, \vec{D}_K) \leqslant 1$$

After a normalization, a global measure of similarity between a user profile P and a document D is computed, based on the combination of the thematic and keyword measures:

$$Sim(P, D) = \delta * Sim_C(\vec{P}_C, \vec{D}_C) + (1 - \delta) * Sim_K(\vec{P}_K, \vec{D}_K)$$

with

$$Sim(P, D) \in \mathbb{Z}, 0 \leqslant Sim(P, D) \leqslant 1$$

The δ value weights the importance given to the thematic similarity compared to the keyword similarity.

This relevance also serves during the ranking process to present ordered relevant sources to the user. Thereby, the use of a search engine on the set of collected Web pages is not necessary. On-topic pages are automatically provided to the user according to their score of relevance without considering their popularity.

3.3 Web source discovery process

The approach we presented includes the modelling of the user profile and a similarity measure used to discover on-topic pages of interest, and to rank relevant collected pages. The steps of the overall process of Web sources discovery implemented in our system called DOWSER (Discovery Of Web Sources Evaluating Relevance) are the following:

1) The user provides URLs of pages according to her needs.
2) DOWSER extracts a list of concepts and a list of keywords from this set of pages to build a user profile.

3) DOWSER adds and rank the URLs into the list of pages to explore.
4) DOWSER collects the first page of the ranked list of pages.
5) The relevance of the collected page according to the user profile is calculated.
6) Outgoing links are extracted.
7) DOWSER adds extracted links to the list of pages to explore with a priority equals to the relevance of the collected page.
8) If the time allocated to the crawl is not over, the process returns at step 4)
9) DOWSER ranks collected pages according to their relevance and top ranked pages are presented to the user.

The time limit at step 8) depends on the aim of the collecting task: in a monitoring system, the time limit of the crawl can be disabled and an alert system replace the two last steps of the process in order to provide relevant pages to the user as soon as they have been collected.

The DOWSER prototype is composed of a modified version of the Heritrix crawler[5] and the processing chain is based on the open source WebLab platform [6].

4. EXPERIMENTAL EVALUATION

In this section, we present three different experiments to evaluate and optimize respectively the user profile construction, the similarity measure and the discovery process.

4.1 User profile evaluation

The first experiment was carried out with 20 experienced users of search engines and IR tools.

To construct her profile, the user provides a set of relevant URLs to the system. The system collects the pages and extracts the text from each page to construct a profile composed of a vector of weighted DBPedia resources and a vector of weighted keywords. A survey is then filled in by the user to assess the relevance of each list of terms considered as a whole. Then, to evaluate the quantity of interesting concepts or keywords automatically extracted, the users mark the terms as *needed*, *related* or *optional*.

Through this experiment, the aim is to validate the two following hypothesis: 1) whether the use of DBPedia resources (thematic) and keywords (specific) is complementary to represent the user's information needs, and 2) whether the automatic extraction is sufficient or if a manual addition of resources by the user is needed.

Results are presented in Figure 2. This result shows that concepts have a better thematic coverage according to the user interests. Concerning the selection of keywords and concepts, the results are presented in Figure 3. The average of selected keywords and concepts was almost the same: 39% of the keywords against 36% of concepts. More globally, the amount of selected needed and related keywords is higher than the amount of selected needed and related concepts. That illustrates how important is the precision, according to the user, during the construction of her profile.

The extraction process of keywords depends on the documents themselves whereas the conceptual extraction is based on a closed conceptual database which does not contain any specific vocabulary. These results show that, in both cases, the automatic construction cannot model the whole user's information needs. The selection of relevant and irrelevant terms by the users validates the use of the semi-automatic extraction method of terms to represent the user needs.

[5]https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

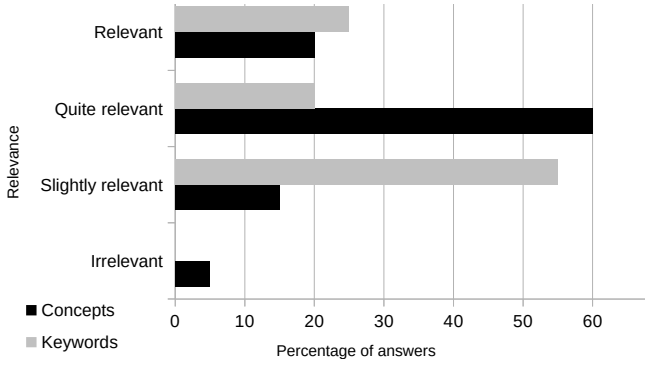

Figure 2: Relevance of the whole lists of extracted concepts and keywords (according to the user)

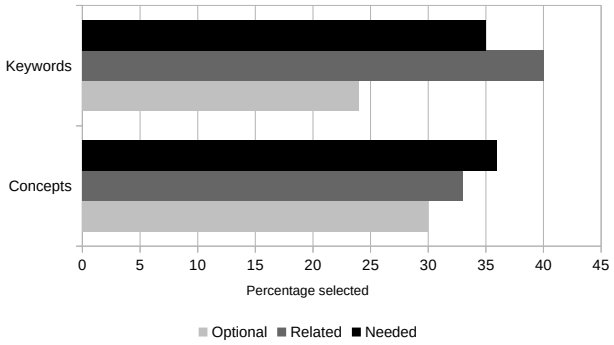

Figure 3: Relevance of extracted terms according to the user feedback

4.2 Similarity measure evaluation

Evaluating the similarity measure of such a discovering system over the whole Web is impossible since all the pages relevant according to a user need are unknown. Thus, the second experiment uses a closed corpus: the *FIRE Collection*[6]. This collection is composed of 15,000 documents and provides questions with a known set of relevant answers from this corpus. We have selected 20 questions and for each only 5 relevant documents, among all the relevant ones, to construct the profile. In this experiment we assess the capability of the system to rank the remaining relevant documents. This experiment is more similar to an IR task than to a DI task since the discovery and exploration processes do not have to be performed. Therefore, at each iteration, DOWSER provides the Top-10 pages according to the similarity measure. These 10 pages are then removed from the remaining documents, for the next iterations.

Our similarity measure depends on three parameters: the size of the keyword vector, the size of the concept vector and finally the δ that enables to combine keywords and concepts. Some empirical tests have been carried out to evaluate the best size for each vector. Firstly, the same size for keyword and concept vectors has been tested simultaneously in graduations of 10. Secondly the keyword vector size has been fixed and the concept vector size was evolving in graduations of 5. And thirdly, the process has been reversed with a fixed size for the concept vector. δ values were tested in

[6]http://www.isical.ac.in/~fire/

graduations of 0.2 to optimize our similarity measure. The F-Measure, based on the precision and the recall, enables us to assess our three parameters.

Figure 4 outlines the variations of the F-Measure according to the size of the concept vector for a fixed size of the keyword vector. We do not provide all the graphics of the experiments but the final results on the FIRE collection show a best size of 5 for the concept vector and 60 for keywords.

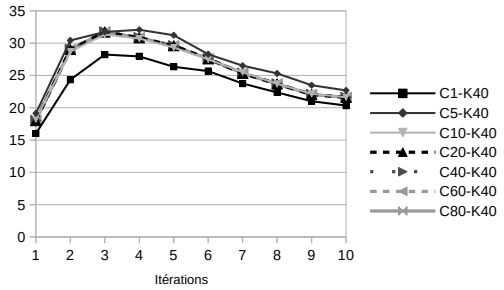

Figure 4: Experiment on vector sizing

Figure 5 illustrates the variation of the F-Measure according to δ. This result shows that combining our two similarity measures into one provides better results than using only the thematic one ($\delta = 1$) or the specific one ($\delta = 0$). According to this figure, the optimized value for δ is 0.6, which means that weighting higher the keyword similarity measure increases favourably the effectiveness of our system.

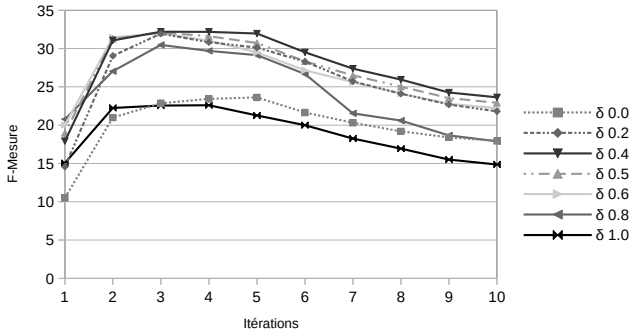

Figure 5: F-Measure according to the δ values

4.3 Discovery process evaluation

The third experiment uses the same process than in section 4.1. Pages collected from the web by DOWSER are presented to the user who evaluates their relevance. Three types of collects are configured. The first one is a focused crawling task using extracted keywords ($\delta = 1$). The second one is a focused crawling task using extracted Concepts ($\delta = 0$). The last one is a traditional breadth-first crawl. Each collected page is annotated with our above similarity measure. The most relevant collected pages from each crawler are shuffled and presented to the user in order to get a feedback: the user labels each presented page as *relevant*, *quite relevant*, *sightly relevant* or *irrelevant*.

Figure 6 shows the result of this evaluation. Among the top-ranked Web pages, 91% of pages provided to users with the keyword-based crawler, and 83% of pages provided with the concept-based crawler, were judged as pages of interest by users according to their information needs. Only 25% of the breadth-first pages were judged interesting.

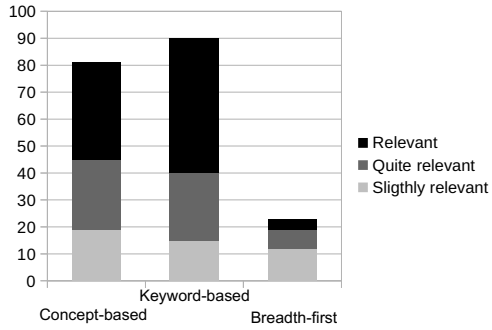

Figure 6: Relevance of provided Web pages

Then, the ability of the DOWSER system to provide the user with relevant Web pages that she cannot find easily using a traditional search engine is evaluated. To this purpose, the user has provided a keyword query representing her information needs, at the end of the construction of the user profile. This query was used at the moment of the collect process to retrieve the top 100 results from Google, Yahoo! and Bing. The domain name of the set of Web pages judged interesting by the user during the feedback task was compared with the domain name of the Google, Yahoo! and Bing top 100. Table 1 synthesises this comparison. 77.4% of the pages provided by the concept-based crawler and 81.5% of the pages provided by the keyword-based crawler had a domain name not present in the Google top 100. Similar results were obtained with Yahoo! and Bing.

Ranked with concepts	Ranked with keywords	
77.4%	81.5%	not in the Google top 100
77.9%	79.8%	not in the Bing top 100
85.7%	82.4%	not in the Yahoo top 100

Table 1: Relevant sources collected with DOWSER not present in the Top 100 of search engines

In order to assess if the pages collected by DOWSER are more relevant than the top pages provided by classic search engines, we finally directly compared the top 5 from DOWSER with the top 5 from Google. Therefore, we provide the user with the list of DOWSER top 5 and Google top 5 pages. The user was asked to score these 10 mixed links between 0 (irrelevant) and 10 (relevant). As presented in Figure 7, the evaluation shows that the average score given by the user to links collected by DOWSER is 6.0 and 4.6 for links from Google. This result is significant according a 95% confidence interval.

These results do not compare the efficiency of the DOWSER prototype with the efficiency of traditional web search engines such as Google. However, they validate the discovery process and illustrate how DOWSER can be used in addition or as an alternative to traditional tools, in order to find new relevant sources according to a user need. The discovery process provides relevant sources regardless

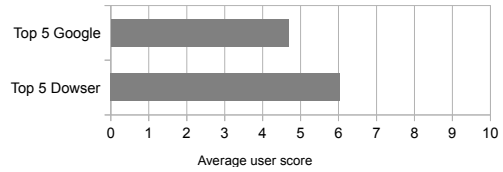

Figure 7: Average user score of the DOWSER and Google Top 5 results

of the popularity of a page [13] which is useful to discover mis-indexed or unpopular Web pages.

5. CONCLUSION AND FUTURE WORK

In this article, we have described a problem faced every day by EIA. Pages on sensitive and specific topics are poorly indexed because of their unpopularity. Search queries and traditional similarity measures do not enable them to find these sources of interest. To tackle this issue, we introduce a bidimensional user profile used by a similarity measure into a focused crawler. The proposed user profile contains a vector of DBPedia concepts and a vector of keywords in order represent the user's need of information both with a thematic and an specific point of view. The similarity measure we introduce is a new relevance calculation based on the bi-dimensional user profile, that leads the exploration of a focused crawler. This approach has been validated within the DOWSER prototype during three experiments. Evaluations show that even if our profile representation cannot model the whole user's information needs, it can be used to assist the crawler in focusing its exploration on relevant Web pages. The weight of the conceptual and terminological parts of the similarity measure has been determined during a second experiment as well as an optimization of the sizes of the two vectors of the user profile. Finally, the discovery process was evaluated and provided results show that our approach can be an alternative to complete search tasks in order to discover relevant pages poorly indexed.

However, when consulting the users of DOWSER, 65% of them note that manually adding terms could improve the representation of their needs. Beyond the construction method, the tools used to extract terms (DBpediaSpotlight [2], NP Chunker [14]) can also have an impact in presented results. Then, the representation of the user needs is limited to the pages provided by the user during the construction of her profile. The next step will be to use relevance feedback on discovered sources to improve the user profile modelling.

To conclude, some real experiments in context with EIA have been carried out but unfortunately, due to the sensitivity of these experiments, results cannot be provided. Informal feedbacks allow to tell that DOWSER fits EIA needs, giving some new interesting informations to EIA.

6. REFERENCES

[1] D. Bergmark, C. Lagoze, and A. Sbityakov. *Research and Advanced Technology for Digital Libraries*, pages 49–70.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.

[3] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[4] E. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497, 2001.

[5] B. Davison. Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279, 2000.

[6] P. Giroux, S. Brunessaux, S. Brunessaux, J. Doucy, G. Dupont, B. Grilheres, Y. Mombrun, and A. Saval. Weblab: An integration infrastructure to ease the development of multimedia processing applications. In *ICSSEA*, 2008.

[7] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[8] J. Kleinberg. *Journal of the ACM (JACM)*, (5):604–632.

[9] F. Liu, C. Yu, and W. Meng. *IEEE Transactions on knowledge and data engineering*, pages 28–40.

[10] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[11] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. *The Adaptive Web*, pages 195–230.

[12] R. Noël, A. Pauchet, B. Grilheres, N. Malandain, L. Vercouter, and S. Brunessaux. Relevant sources of information are not necessarily popular ones. In *WI*, Warsaw, Pologne, 2014.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[14] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94, 1995.

[15] J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Machine Learning - International Workshop THEN Conference-*, pages 335–343, 1999.

[16] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[17] A. Saint Requier, G. Dupont, S. Adam, Y. Lecourtier, and S. Brunessaux. Selection adaptative de services de recherche d'information web en fonction du besoin de l'utilisateur. In *Proceedings of the sos-dlwd2012*, 2012.

[18] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[19] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723, 2006.

[20] A. Wærn. *User Modeling and User-Adapted Interaction*, (2):201–237.

[21] R. White, I. Ruthven, and J. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–42, 2005.

[22] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.